RegEx

Fahad Ausaf[1], Roy Dyckhoff[2], and Christian Urban[1]

[1] King's College London, United Kingdom
[2] St Andrews

Abstract. BLA BLA Sulzmann and Lu [1]

Keywords:

1 Introduction

Regular exprtessions

$$r := \mathit{NULL} \mid \mathit{EMPTY} \mid \mathit{CHAR}\ c \mid \mathit{ALT}\ r_1\ r_2 \mid \mathit{SEQ}\ r_1\ r_2 \mid \mathit{STAR}\ r$$

Values

$$v := \mathit{Void} \mid \mathit{Char}\ c \mid \mathit{Left}\ v \mid \mathit{Right}\ v \mid \mathit{Seq}\ v_1\ v_2 \mid \mathit{Stars}\ vs$$

The language of a regular expression

$$\begin{array}{lcl}
L\ \mathit{NULL} & \overset{\text{def}}{=} & \varnothing \\
L\ \mathit{EMPTY} & \overset{\text{def}}{=} & \{[]\} \\
L\ (\mathit{CHAR}\ c) & \overset{\text{def}}{=} & \{[c]\} \\
L\ (\mathit{SEQ}\ r_1\ r_2) & \overset{\text{def}}{=} & (L\ r_1)\ @\ (L\ r_2) \\
L\ (\mathit{ALT}\ r_1\ r_2) & \overset{\text{def}}{=} & (L\ r_1) \cup (L\ r_2) \\
L\ (\mathit{STAR}\ r) & \overset{\text{def}}{=} & (L\ r)\star
\end{array}$$

The nullable function

$$\begin{array}{lcl}
\mathit{nullable}\ \mathit{NULL} & \overset{\text{def}}{=} & \mathit{False} \\
\mathit{nullable}\ \mathit{EMPTY} & \overset{\text{def}}{=} & \mathit{True} \\
\mathit{nullable}\ (\mathit{CHAR}\ c) & \overset{\text{def}}{=} & \mathit{False} \\
\mathit{nullable}\ (\mathit{ALT}\ r_1\ r_2) & \overset{\text{def}}{=} & \mathit{nullable}\ r_1 \vee \mathit{nullable}\ r_2 \\
\mathit{nullable}\ (\mathit{SEQ}\ r_1\ r_2) & \overset{\text{def}}{=} & \mathit{nullable}\ r_1 \wedge \mathit{nullable}\ r_2 \\
\mathit{nullable}\ (\mathit{STAR}\ r) & \overset{\text{def}}{=} & \mathit{True}
\end{array}$$

The derivative function for characters and strings

$$
\begin{array}{lcl}
\mathit{der}\;c\;\mathit{NULL} & \stackrel{\text{def}}{=} & \mathit{NULL}\\
\mathit{der}\;c\;\mathit{EMPTY} & \stackrel{\text{def}}{=} & \mathit{NULL}\\
\mathit{der}\;c\;(\mathit{CHAR}\;c') & \stackrel{\text{def}}{=} & \mathit{if}\;c = c'\;\mathit{then}\;\mathit{EMPTY}\;\mathit{else}\;\mathit{NULL}\\
\mathit{der}\;c\;(\mathit{ALT}\;r_1\;r_2) & \stackrel{\text{def}}{=} & \mathit{ALT}\;(\mathit{der}\;c\;r_1)\;(\mathit{der}\;c\;r_2)\\
\mathit{der}\;c\;(\mathit{SEQ}\;r_1\;r_2) & \stackrel{\text{def}}{=} & \mathit{if}\;\mathit{nullable}\;r_1\;\mathit{then}\;\mathit{ALT}\;(\mathit{SEQ}\;(\mathit{der}\;c\;r_1)\;r_2)\;(\mathit{der}\;c\;r_2)\\
 & & \mathit{else}\;\mathit{SEQ}\;(\mathit{der}\;c\;r_1)\;r_2\\
\mathit{der}\;c\;(\mathit{STAR}\;r) & \stackrel{\text{def}}{=} & \mathit{SEQ}\;(\mathit{der}\;c\;r)\;(\mathit{STAR}\;r)\\
\\
\mathit{ders}\;[]\;r & \stackrel{\text{def}}{=} & r\\
\mathit{ders}\;(c{::}s)\;r & \stackrel{\text{def}}{=} & \mathit{ders}\;s\;(\mathit{der}\;c\;r)
\end{array}
$$

The *flat* function for values

$$
\begin{array}{lcl}
|\mathit{Void}| & \stackrel{\text{def}}{=} & []\\
|\mathit{Char}\;c| & \stackrel{\text{def}}{=} & [c]\\
|\mathit{Left}\;v| & \stackrel{\text{def}}{=} & |v|\\
|\mathit{Right}\;v| & \stackrel{\text{def}}{=} & |v|\\
|\mathit{Seq}\;v_1\;v_2| & \stackrel{\text{def}}{=} & |v_1|\;@\;|v_2|\\
|\mathit{Stars}\;[]| & \stackrel{\text{def}}{=} & []\\
|\mathit{Stars}\;(v{::}vs)| & \stackrel{\text{def}}{=} & |v|\;@\;|\mathit{Stars}\;vs|
\end{array}
$$

The *mkeps* function

$$
\begin{array}{lcl}
\mathit{mkeps}\;\mathit{EMPTY} & \stackrel{\text{def}}{=} & \mathit{Void}\\
\mathit{mkeps}\;(\mathit{SEQ}\;r_1\;r_2) & \stackrel{\text{def}}{=} & \mathit{Seq}\;(\mathit{mkeps}\;r_1)\;(\mathit{mkeps}\;r_2)\\
\mathit{mkeps}\;(\mathit{ALT}\;r_1\;r_2) & \stackrel{\text{def}}{=} & \mathit{if}\;\mathit{nullable}\;r_1\;\mathit{then}\;\mathit{Left}\;(\mathit{mkeps}\;r_1)\;\mathit{else}\;\mathit{Right}\;(\mathit{mkeps}\;r_2)\\
\mathit{mkeps}\;(\mathit{STAR}\;r) & \stackrel{\text{def}}{=} & \mathit{Stars}\;[]
\end{array}
$$

The *inj* function

$$
\begin{array}{lcl}
\mathit{inj}\;\mathit{EMPTY}\;c\;\mathit{Void} & \stackrel{\text{def}}{=} & \mathit{Char}\;c\\
\mathit{inj}\;(\mathit{CHAR}\;d)\;c\;\mathit{Void} & \stackrel{\text{def}}{=} & \mathit{Char}\;d\\
\mathit{inj}\;(\mathit{CHAR}\;d)\;c\;(\mathit{Char}\;c') & \stackrel{\text{def}}{=} & \mathit{Seq}\;(\mathit{Char}\;d)\;(\mathit{Char}\;c')\\
\mathit{inj}\;(\mathit{ALT}\;r_1\;r_2)\;c\;(\mathit{Left}\;v_1) & \stackrel{\text{def}}{=} & \mathit{Left}\;(\mathit{inj}\;r_1\;c\;v_1)\\
\mathit{inj}\;(\mathit{ALT}\;r_1\;r_2)\;c\;(\mathit{Right}\;v_2) & \stackrel{\text{def}}{=} & \mathit{Right}\;(\mathit{inj}\;r_2\;c\;v_2)\\
\mathit{inj}\;(\mathit{SEQ}\;r_1\;r_2)\;c\;(\mathit{Seq}\;v_1\;v_2) & \stackrel{\text{def}}{=} & \mathit{Seq}\;(\mathit{inj}\;r_1\;c\;v_1)\;v_2\\
\mathit{inj}\;(\mathit{SEQ}\;r_1\;r_2)\;c\;(\mathit{Left}\;(\mathit{Seq}\;v_1\;v_2)) & \stackrel{\text{def}}{=} & \mathit{Seq}\;(\mathit{inj}\;r_1\;c\;v_1)\;v_2\\
\mathit{inj}\;(\mathit{SEQ}\;r_1\;r_2)\;c\;(\mathit{Right}\;v_2) & \stackrel{\text{def}}{=} & \mathit{Seq}\;(\mathit{mkeps}\;r_1)\;(\mathit{inj}\;r_2\;c\;v_2)\\
\mathit{inj}\;(\mathit{STAR}\;r)\;c\;(\mathit{Seq}\;v\;(\mathit{Stars}\;vs)) & \stackrel{\text{def}}{=} & \mathit{Stars}\;((\mathit{inj}\;r\;c\;v){::}vs)
\end{array}
$$

The inhabitation relation:

$$\frac{\vdash v_1 : r_1 \qquad \vdash v_2 : r_2}{\vdash \mathit{Seq}\ v_1\ v_2 : \mathit{SEQ}\ r_1\ r_2}$$

$$\frac{\vdash v_1 : r_1}{\vdash (\mathit{Left}\ v_1) : \mathit{ALT}\ r_1\ r_2} \qquad \frac{\vdash v_2 : r_1}{\vdash (\mathit{Right}\ v_2) : \mathit{ALT}\ r_2\ r_1}$$

$$\frac{}{\vdash \mathit{Void} : \mathit{EMPTY}} \qquad \frac{}{\vdash (\mathit{Char}\ c) : \mathit{CHAR}\ c}$$

$$\frac{}{\vdash \mathit{Stars}\ [] : \mathit{STAR}\ r} \qquad \frac{\vdash v : r \qquad \vdash \mathit{Stars}\ \mathit{vs} : \mathit{STAR}\ r}{\vdash \mathit{Stars}\ (v{::}\mathit{vs}) : \mathit{STAR}\ r}$$

We have also introduced a slightly restricted version of this relation where the last rule is restricted so that $|v| \neq []$. This relation for *non-problematic* is written $\models v : r$.

Our Posix relation $s \in r \to v$

$$\frac{}{[] \in \mathit{EMPTY} \to \mathit{Void}} \qquad \frac{}{[c] \in \mathit{CHAR}\ c \to (\mathit{Char}\ c)}$$

$$\frac{s \in r_1 \to v}{s \in \mathit{ALT}\ r_1\ r_2 \to (\mathit{Left}\ v)} \qquad \frac{s \in r_2 \to v \qquad s \notin (L\ r_1)}{s \in \mathit{ALT}\ r_1\ r_2 \to (\mathit{Right}\ v)}$$

$$\frac{\begin{array}{c} s_1 \in r_1 \to v_1 \qquad s_2 \in r_2 \to v_2 \\ \nexists s_3\ s_4.\ s_3 \neq [] \wedge s_3 \mathbin{@} s_4 = s_2 \wedge s_1 \mathbin{@} s_3 \in (L\ r_1) \wedge s_4 \in (L\ r_2) \end{array}}{(s_1 \mathbin{@} s_2) \in \mathit{SEQ}\ r_1\ r_2 \to \mathit{Seq}\ v_1\ v_2}$$

$$\frac{\begin{array}{c} s_1 \in r \to v \qquad s_2 \in \mathit{STAR}\ r \to \mathit{Stars}\ \mathit{vs} \\ |v| \neq [] \qquad \nexists s_3\ s_4.\ s_3 \neq [] \wedge s_3 \mathbin{@} s_4 = s_2 \wedge s_1 \mathbin{@} s_3 \in (L\ r) \wedge s_4 \in (L\ (\mathit{STAR}\ r)) \end{array}}{(s_1 \mathbin{@} s_2) \in \mathit{STAR}\ r \to \mathit{Stars}\ (v{::}\mathit{vs})}$$

$$\frac{}{[] \in \mathit{STAR}\ r \to \mathit{Stars}\ []}$$

Our version of Sulzmann's ordering relation

$$\frac{v_1 \succeq_{r_1} {v_1}' \qquad v_1 \neq {v_1}'}{\mathit{Seq}\ v_1\ v_2 \succeq_{\mathit{SEQ}\ r_1\ r_2} \mathit{Seq}\ {v_1}'\ {v_2}'} \qquad \frac{v_2 \succeq_{r_2} {v_2}'}{\mathit{Seq}\ v_1\ v_2 \succeq_{\mathit{SEQ}\ r_1\ r_2} \mathit{Seq}\ v_1\ {v_2}'}$$

$$\frac{\mathit{len}\ (|v_1|) \leq \mathit{len}\ (|v_2|)}{\mathit{Left}\ v_2 \succeq_{\mathit{ALT}\ r_1\ r_2} \mathit{Right}\ v_1} \qquad \frac{\mathit{len}\ (|v_2|) < \mathit{len}\ (|v_1|)}{\mathit{Right}\ v_1 \succeq_{\mathit{ALT}\ r_1\ r_2} \mathit{Left}\ v_2}$$

$$\frac{v_2 \succeq_{r_2} {v_2}'}{\mathit{Right}\ v_2 \succeq_{\mathit{ALT}\ r_1\ r_2} \mathit{Right}\ {v_2}'} \qquad \frac{v_1 \succeq_{r_1} {v_1}'}{\mathit{Left}\ v_1 \succeq_{\mathit{ALT}\ r_1\ r_2} \mathit{Left}\ {v_1}'}$$

$$\frac{}{\mathit{Void} \succeq_{\mathit{EMPTY}} \mathit{Void}} \qquad \frac{}{\mathit{Char}\ c \succeq_{\mathit{CHAR}\ c} \mathit{Char}\ c}$$

$$\frac{|\mathit{Stars}\ (v{::}vs)| = []}{\mathit{Stars}\ [] \succeq_{\mathit{STAR}\ r} \mathit{Stars}\ (v{::}vs)} \qquad \frac{|\mathit{Stars}\ (v{::}vs)| \neq []}{\mathit{Stars}\ (v{::}vs) \succeq_{\mathit{STAR}\ r} \mathit{Stars}\ []}$$

$$\frac{v_1 \succeq_r v_2}{\mathit{Stars}\ (v_1{::}vs_1) \succeq_{\mathit{STAR}\ r} \mathit{Stars}\ (v_2{::}vs_2)}$$

$$\frac{\mathit{Stars}\ vs_1 \succeq_{\mathit{STAR}\ r} \mathit{Stars}\ vs_2}{\mathit{Stars}\ (v{::}vs_1) \succeq_{\mathit{STAR}\ r} \mathit{Stars}\ (v{::}vs_2)} \qquad \frac{}{\mathit{Stars}\ [] \succeq_{\mathit{STAR}\ r} \mathit{Stars}\ []}$$

A prefix of a string s

$$s_1 \sqsubseteq s_2 \overset{\mathit{def}}{=} \exists\, s3.\ s_1 \mathbin{@} s3 = s_2$$

Values and non-problematic values

$$\mathit{Values}\ r\ s \overset{\mathit{def}}{=} \{v \mid \vdash v : r \wedge (|v|) \sqsubseteq s\}$$

$$\mathit{NValues}\ r\ s \overset{\mathit{def}}{=} \{v \mid \models v : r \wedge (|v|) \sqsubseteq s\}$$

The point is that for a given s and r there are only finitely many non-problematic values.

Some lemmas we have proved:

$(L\ r) = \{|v| \mid \vdash v : r\}$
$(L\ r) = \{|v| \mid \models v : r\}$
If $\mathit{nullable}\ r$ *then* $\vdash \mathit{mkeps}\ r : r$.
If $\mathit{nullable}\ r$ *then* $|\mathit{mkeps}\ r| = []$.
If $\vdash v : \mathit{der}\ c\ r$ *then* $\vdash (\mathit{inj}\ r\ c\ v) : r$.
If $\vdash v : \mathit{der}\ c\ r$ *then* $|\mathit{inj}\ r\ c\ v| = c{::}(|v|)$.
If $\mathit{nullable}\ r$ *then* $[] \in r \to \mathit{mkeps}\ r$.
If $s \in r \to v$ *then* $|v| = s$.
If $s \in r \to v$ *then* $\models v : r$.
If $s \in r \to v_1$ *and* $s \in r \to v_2$ *then* $v_1 = v_2$.

This is the main theorem that lets us prove that the algorithm is correct according to $s \in r \to v$:
If $s \in \mathit{der}\ c\ r \to v$ *then* $(c{::}s) \in r \to (\mathit{inj}\ r\ c\ v)$.

Things we have proved about our version of the Sulzmann ordering

$$\textit{If } \vdash v : r \textit{ then } v \succeq_r v.$$

$$\textit{If } \vdash v_1 : r \textit{ and } \vdash v_2 : r \textit{ then } v_1 \succeq_r v_2 \lor v_2 \succeq_r v_1.$$

Things we like to prove, but cannot:

If $s \in r \to v_1, \vdash v_2 : r$, then $v_1 \succeq_r v_2$

References

1. M. Sulzmann and K. Lu. POSIX Regular Expression Parsing with Derivatives. In *Proc. of the 12th International Conference on Functional and Logic Programming (FLOPS)*, volume 8475 of *LNCS*, pages 203–220, 2014.